32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

An information theory-based tool for characterizing the interaction environment of a protein.

Raimon Massanet-Vila[1,2,3], Joan-Josep Gallardo-Chacón[3], Pere Caminal[1,2,3], Alexandre Perera[1,2,3]

[1]Dept. ESAII, Technical University of Catalonia (UPC), Barcelona, Spain.
Email: {raimon.massanet, pere.caminal, alexandre.perera}@upc.edu

[2]Biomedical Engineering Research Center (CREB), Barcelona, Spain.

[3]CIBER-BBN in Bioengineering, Biomaterials and Nanomedicine, Spain.
Email: joan.josep.gallardo@upc.edu

Abstract— **In recent years large amounts of information have been accumulated in proteomic, genetic and metabolic databases. Much effort has been dedicated to developing methods that successfully exploit, organize and structure this information. However, there is no application, that we know of, that semantically characterizes the interaction environment in which a protein exists. A high-throughput software package has been developed to retrieve information from publicly available databases, such as the Gene Ontology Annotation (GOA) database and the Human Proteome Resource Database (HPRD) and structure their information. This information is presented to the user as groups of semantically described dense interaction subnetworks that interact with a target protein.**

I. INTRODUCTION

Most of the biological processes need combined and synchronized activity of protein sets forming metabolic, signaling and regulatory pathways in cells [1]. The experimental data about proteins and their role in organisms' functioning has been compiled and organized in large databases which are very useful to extract biological relationships [2]. However, the information about biological systems needs to be organized and prioritized to attend the special characteristics of each research performance. To solve this practical problems, different bioinformatics tools are appearing and are being evaluated [3], [4]. Ontologies, like the Gene Ontology, have been defined in order to standardize semantic information that many different research groups over the world discover about proteins [5]. At the same time, they allow for semantic information to be retrieved, processed and even generated by computer programs.

The great amount of data accumulated over the last years could be crucial in further developing genetics, proteomics and metabolomics. Understanding cell processes and the gene mutations that disrupt them is a major goal in these disciplines. However, retrieving and exploiting large amounts of information can be very challenging. Clustering is a well-known and widely accepted technique for exploratory data analysis. Partitioning a large set of data in smaller and more compact groups helps in understanding the structure of the data. Spectral clustering is a clustering technique that has recently become very popular [6], [7]. It has been shown that some implementations of this technique solve the problem of partitioning a graph from different points of view at the same time: graph cut, random walk and perturbation theory [8].

By means of combining protein-protein interaction, spectral clustering and the Gene Ontology it is possible to enrich the information of protein networks [9] and extract valuable information about the different clusters in order to explain the relations between proteins. Once this large amount of information is organized could be very useful for studying macroscopic problems such as diseases or physiological models.

This work proposes a methodology for automatically extracting information from publicly available databases and organize it so that useful knowledge can be extracted about the local interaction environment of a protein. This is accomplished by partitioning the environment in dense interaction subnetworks and find the semantic labels in each subnetwork that best describe it. This could be useful to researchers that are beginning to study a protein and need to have background on it. A software package has been written in the R statistical programming language that implements the methodology [10]. This software package characterizes the processes in which the protein and its immediate interacting neighbors participate.

II. MATERIALS AND METHODS

A. Methodology

In order to characterize the interaction environment of a protein we propose a methodology based on partitioning its local interaction domain [11]. The idea is to find dense subnetworks in this domain and characterize them using semantic annotations from the Gene Ontology. Statistical

This work was supported by the Spanish Ministerio de Educación y Ciencia under the Ramón y Cajal Program and TEC2007-63637/TCM and by the ISCIII under the CIBER initiative.

tests are performed between the dense subnetworks in order to find the set of annotations that best describes them. The methodology was divided in four steps, detailed below. Throughout this text, the term *interaction* refers to physical binary interactions (docking) between two proteins.

1) Interaction mining: In the first step we used the Human Proteome Resource Database (HPRD) [12] to build a local interaction environment for the protein under study. This environment contains all proteins that are less than n interactions away from the protein. This parameter n was defined in order to control the amount of environment information to be taken into account. This information was retrieved from the HPRD database and modeled as an interaction network using an undirected graph structure, where nodes represent proteins and edges represent interactions between proteins.

2) Clustering: The graph representing the protein's interaction environment was partitioned using spectral clustering. Spectral clustering is a technique that uses a graph's Laplacian matrix to partition the graph.

This is implemented by the software package as follows. Let D be the degree matrix so that d_{ii} is the degree of node i and $d_{ij} = 0$ for $i \neq j$. Let W be the weight matrix so that w_{ij} is a measure of similarity between nodes i and j. In this work we used the graph's adjacency matrix as weight matrix:

$$w_{ij} = \begin{cases} 1 & \text{if proteins } i \text{ and } j \text{ interact} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

However, other weight matrices could have been used (number of publications linking the two proteins, semantic similarity, etc). Three different Laplacian matrices have been implemented to create a spectral model, following [8]. The unnormalized Laplacian matrix is defined as:

$$L_u = D - W \quad (2)$$

The symmetric normalized Laplacian matrix:

$$L_{sym} = I - D^{-1/2}\, W\, D^{-1/2} \quad (3)$$

And the random-walk normalized Laplacian matrix:

$$L_{rw} = I - D^{-1}\, W \quad (4)$$

The Laplacian matrix to be used can be chosen by the user.

Once the spectral model is built, the user is presented a plot of the eigenvalues and asked for the number of clusters k to partition the data. Then the spectral model is used to partition the interaction network in k clusters.

3) Semantic mining: The Gene Ontology Annotation (GOA) [13] database was used to enrich the partitioned interaction network with semantic information regarding *biological process*. For each protein a set of semantic annotations were retrieved. This yielded a distinct distribution of semantic annotations within each cluster.

At this step the software gives the user the opportunity to exclude annotations that have a certain evidence code — usually electronically inferred annotations (IEA) are discarded.

4) Statistics: Finally, statistical tests were performed in order to find annotations that were characteristic of each cluster. To do so, the annotations of every cluster were compared with a null distribution. A null distribution was generated for each annotation of each cluster as the distribution of the annotation in 50 random samples of the same size as the cluster. Random samples were taken from proteins belonging to the interaction environment. Empirical p-values were calculated by modeling the function of the null distribution using a kernel approach [14]. This yielded a level of significance of each annotation inside each cluster, indicating how different was the distribution of the annotation in that particular cluster from the same annotation in the environment. By selecting the annotations that were most differentially found in every cluster the user can obtain a semantic description of every cluster. This allows for a modular semantic description of the environment a protein interacts with.

B. Case study

In order to test the methodology a case study was proposed by a biological expert. The method was applied to human small heat shock protein (HSP27) in order to study its environment. The parameter n was set to 3 because this value yielded a network with a good compromise between amount of information and handleable size. A random-walk Laplacian matrix (L_{rw}) was used to build the spectral model. The interaction network was partitioned into 8 clusters after an examination of the spectral model's eigenvalues because this value showed a considerable increase in the eigenvalues' derivative. After the network was partitioned and semantically enriched, the statistical tests were performed. For each cluster, the number of significantly differentially distributed annotations was compared with the total number of annotations. The results of this case study are presented in the next section III.

III. RESULTS

The interaction environment obtained for protein HSP27 consisted of 601 proteins and 2840 interactions. Fig. 1 shows the network's node distribution, which is characterized by very few highly-connected proteins, and many proteins with low connectivity. 7 of the 8 partitions had at least one significantly differentially distributed semantic label that allowed for an interpretation of the cluster. The other cluster contained only one semantic annotation that was not found to be significant. Fig. 2 shows the number of proteins in each cluster. Clusters 3, 7 and 8 contain most of the proteins of the interaction environment.

Cluster 3 contains the initial protein, HSP27, and the majority of the rest of the proteins. This large dense subnetwork is typical of scale-free networks and indicates that HPS27 has a role in a very tightly connected set of biological functions. This group of proteins is semantically characterized by the highly significantly differentially distributed annotations. Table I shows the 10 most statistically significant semantic labels describing biological processes. The results suggest that this large group of proteins has a major regulating role

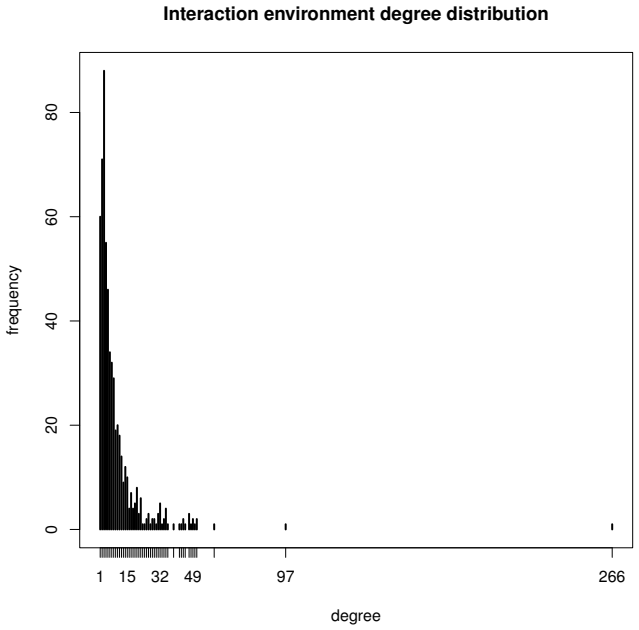

Fig. 1: Degree distribution of the interaction environment network for protein HSP27 with a maximum of 3 levels of interaction.

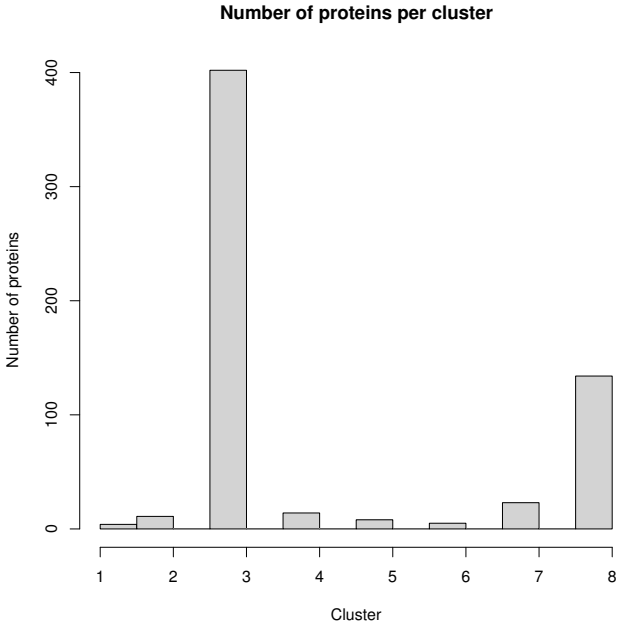

Fig. 2: Number of proteins in each cluster after applying spectral clustering.

in several processes, including factor-beta2 production, nitric oxide biosynthetic process and apoptosis.

Other large dense subnetworks have been found in clusters 8 and 7. This result suggests and indirect relationship of protein HSP27 with other biological processes. Biological processes in cluster 8 are best described by the annotations found in Table III, which suggest that it contains proteins involved in response to stimulus, virus and correction of transcription errors. On the other hand, cluster 7 is described in Table II as a group of proteins highly involved in protein transport and cell motion.

For space limitations, only the 10 most significant labels were presented, and similar tables for the rest of the clusters were omitted. However, the user has the opportunity of exploring the semantic description of all the clusters interactively.

TABLE I: Semantic description of cluster 3

	label	pvalue	dir
1	activation of JNK activity	4.23e-19	high
2	DNA ligation	5.65e-19	high
3	regulation of transforming growth factor-beta2 production	5.65e-19	high
4	positive regulation of nitric oxide biosynthetic process	6.52e-19	high
5	negative regulation of protein kinase activity	1.27e-18	high
6	nitric oxide biosynthetic process	1.28e-18	high
7	positive regulation of protein binding	1.28e-18	high
8	positive regulation of transcription	1.35e-18	high
9	base-excision repair	1.35e-18	high
10	induction of apoptosis	1.47e-18	high

TABLE II: Semantic description of cluster 7

	label	pvalue	dir
1	actin filament bundle formation	1.34e-18	high
2	actin cytoskeleton organization and biogenesis	4.96e-18	high
3	cell motility	1.26e-17	high
4	ovarian follicle development	9.74e-03	high
5	oocyte maturation	9.74e-03	high
6	diacylglycerol biosynthetic process	9.74e-03	high
7	calcium ion-dependent exocytosis	9.74e-03	high
8	actin filament polymerization	9.74e-03	high
9	actin filament-based movement	9.74e-03	high
10	actomyosin structure organization and biogenesis	9.74e-03	high

Table IV shows the number of significant and non-significant labels in each cluster. It shows that all clusters have an important proportion of significant labels. This result suggests that the network partitioning obtained by using only the graph structure is also semantically coherent.

IV. DISCUSSION

This work offers a fast method to semantically characterize the interaction environment of a protein. A case study has been presented as a demonstration of the methodology. The

TABLE III: Semantic description of cluster 8

	label	pvalue	dir
1	postreplication repair	8.15e-18	high
2	RNA processing	7.36e-03	high
3	nuclear mRNA splicing, via spliceosome	7.54e-03	high
4	alcohol metabolic process	9.74e-03	high
5	inosine catabolic process	9.74e-03	high
6	unknown GO label	9.74e-03	high
7	segment specification	9.74e-03	high
8	response to biotic stimulus	9.74e-03	high
9	response to virus	1.58e-02	high
10	response to unfolded protein	1.64e-02	high

TABLE IV: Number of labels in each cluster

	labels	significant labels
1	5	1
2	10	4
3	666	337
4	20	9
5	16	8
6	1	0
7	59	31
8	131	33

results show that the partition defines groups of proteins with a high semantic coherence and meaningful labels. The method is a high-throughput and automatic tool to retrieve information from publicly available databases and present a description to the user using widely accepted semantic labels.

V. ACKNOWLEDGMENTS

The authors want to acknowledge the support received from the Spanish Ministerio de Educación y Ciencia under the Ramón y Cajal Program and TEC2007-63637/TCM and by the Instituto de Salud Carlos III under the initiative CIBER-BBN in Bioengineering, Biomaterials and Nanomedicine.

The authors want to thank the reviewers for their comments.

REFERENCES

[1] M. Monti, M. Cozzolino, F. Cozzolino, G. Vitiello, R. Tedesco, A. Flagiello, and P. Pucci, "Puzzle of protein complexes in vivo: a present and future challenge for functional proteomics." *Expert Rev Proteomics*, vol. 6, no. 2, pp. 159–169, Apr 2009. [Online]. Available: 10.1586/epr.09.7; http://dx.doi.org/10.1586/epr.09.7

[2] M. Krallinger, F. Leitner, and A. Valencia, "Analysis of biological processes and diseases using text mining approaches." *Methods Mol Biol*, vol. 593, pp. 341–382, 2010.

[3] A. A. Terentiev, N. T. Moldogazieva, and K. V. Shaitan, "Dynamic proteomics in modeling of the living cell. protein-protein interactions." *Biochemistry (Mosc)*, vol. 74, no. 13, pp. 1586–1607, Dec 2009.

[4] K. Raman, "Construction and analysis of protein-protein interaction networks." *Autom Exp*, vol. 2, no. 1, p. 2, 2010. [Online]. Available: 10.1186/1759-4499-2-2; http://dx.doi.org/10.1186/1759-4499-2-2

[5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium," *Nature genetics*, vol. 25, no. 1, pp. 25–29, May 2000, lR: 20071115; GR: HD33745/HD/NICHD NIH HHS/United States; GR: P41 HG00330/HG/NHGRI NIH HHS/United States; GR: P41 HG01315/HG/NHGRI NIH HHS/United States; GR: etc.; JID: 9216904; ppublish.

[6] W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs," *IBM J.Res.Dev.*, vol. 17, no. 5, pp. 420–425, 1973.

[7] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Mathematical Journal*, vol. 23, no. 2, pp. 298–305, 1973.

[8] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, December 2007. [Online]. Available: http://dx.doi.org/10.1007/s11222-007-9033-z; http://dx.doi.org/10.1007/s11222-007-9033-z

[9] M. Deng, Z. Tu, F. Sun, and T. Chen, "Mapping gene ontology to proteins based on protein-protein interaction data." *Bioinformatics*, vol. 20, no. 6, pp. 895–902, Apr 2004. [Online]. Available: 10.1093/bioinformatics/btg500; http://dx.doi.org/10.1093/bioinformatics/btg500

[10] R Development Core Team, "R: A language and environment for statistical computing," 2009. [Online]. Available: http://www.R-project.org

[11] R. Massanet-Vila, J. J. Gallardo-Chacón, P. Caminal, and A. Perera, "Search of phenotype related candidate genes using gene ontology-based semantic similarity and protein interaction information: Application to brugada syndrome," in *31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC2009*, 2009.

[12] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human protein reference database–2009 update," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D767–772, January 1 2009.

[13] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler, "The goa database in 2009–an integrated gene ontology annotation resource," *Nucl.Acids Res.*, p. gkn803, October 2008. [Online]. Available: 10.1093/nar/gkn803; http://dx.doi.org/10.1093/nar/gkn803

[14] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society.Series B (Methodological)*, vol. 53, no. 3, pp. 683–690, 1991. [Online]. Available: http://www.jstor.org/stable/2345597